云 南 植 物 研 究 2010，32(5)：401～406
Acta Botanica Yunnanica DOI：10.3724/SP.J.1143.2010.10101

中国植物名称数据库的建设及若干问题的探讨*

张 宇[1,2]，王雨华[1**]

（1 中国科学院昆明植物研究所资源植物与生物技术重点实验室，云南 昆明 650204；
2 中国科学院研究生院，北京 100049）

摘要：中国植物物种信息数据库是中国科学院科学数据库参考型数据库，中国植物名称数据库（CPNI）是中国植物物种信息数据库最重要的组成部分，是中国植物物种信息数据库收录植物的目录和索引，也是其他数据库的参考和联系的桥梁。以《中国植物志》和《Flora of China》为基础数据来源，加上少量参考数据作为补充，设计建设了中国植物名称数据库，并从现有已建成同类数据库的评价、数据来源和数据组成，以及建库策略实现对中国植物名称数据库建设进行了分析和探讨。中国植物名称数据库是植物名称的参考型数据库，能够辅助植物学相关学科研究中关于植物名称的研究和利用。

关键词：CPNI；植物名称；科学数据库；数据标准
中图分类号：Q 949，G 350 **文献标识码**：A **文章编号**：0253-2700(2010) 05-401-06

The Researching and Discussion on the Construction of the Database of Chinese Plant Names Index (CPNI)*

ZHANG Yu[1,2], WANG Yu-Hua[1**]

(1 *Key Laboratory of Economic Plants and Biotechnology, Kunming Institute of Botany, Chinese Academy of Sciences*, Kunming 650204, China; 2 *Graduate University of Chinese Academy of Sciences*, Beijing 100049, China)

Abstract: The Chinese Plants reference database (CPRD) is a part of the Scientific Database. The Chinese Plant Names Index (CPNI) is one of the important parts of CPRD, because it is the names index of CPRD and the bridge to other plant databases. Based on the Chinese version and English version of "Flora of China", the CPNI was constructed with other little reference data. Then, the advantages and disadvantages of the databases of plant names which have been published in the world, the data resources and the construction strategies of the Chinese plant names was discussed. The CPNI is a reference database of the Chinese Plant Names which can be used on studies on plant sciences especially when the names of the plants which are studied on must be ensured.

Key words: CPNI; Plant name; Scientific database; Data standard

我国是世界上 10 个生物多样性大国之一，生物物种极为丰富，中国拥有高等植物 3 万余种，其中苔藓、蕨类、裸子植物和被子植物分别占世界总数的 9.1%、22%、26.7%和 10%（裴盛基等，2009）。名称是一个事物区别与另一个事物的直接标识。而植物的“名称标识”就是植物的唯一合法学名，且每一个分类等级都具有唯一的合法学名（张丽兵译，2007）。对于植物学研究者来说，确定研究对象的唯一合法学名是整个研究的首要前提。中国植物名称数据库（the

* 基金项目：中科院科学数据库项目（INFO-115-C01-SDB1-02）；科技部科技基础性工作专项重点项目（2007FY110100）；云南省科技计划项目（2009CC016）
** 通讯作者：Author for correspondence；E-mail：wangyuhua@mail.kib.ac.cn
收稿日期：2010-05-10，2010-07-20 接受发表
作者简介：张宇（1985－）男，在读硕士研究生，主要从事植物学，植物资源与生物多样性信息方向研究工作。

Chinese Plant Names Index，CPNI）是依照中国科学院科学数据库中参考型数据库建设标准规范（中国科学院计算机网络信息中心科学数据中心，2009），以中国科学院昆明植物研究所为主的相关研究机构长期积累的数据为基础建成的符合国家或国际标准、有严格质量控制与管理、内容具有完整性和权威性的数据库。它是《中国植物物种信息数据库》的一部分，也是整个物种信息数据库的物种名录和查询索引；是连接物种信息库中各个子数据库的桥梁，也是物种数据库和其他数据库之间联系的纽带。所以成功建成中国植物名称数据库是整个中国植物物种信息数据库建设成功的关键所在，在建设过程中产生的相关问题和结论亦值得深入研究探讨。

1 国内外同类数据库的建设情况及其评价

自 1992 年环境与发展大会召开以来，各国政府和从事生物多样性保护的国际组织普遍提高了对生物多样性信息收集和管理的重视程度，开发建立了大量的联网数据库和网站。比较知名的有 species2000，国际植物名录（The International Plant Names Index，IPNI）、the Integrated Taxonomic Information System（ITIS）、International Organization for Plant Information（IOPI）、TROPICOS® 等等（Singh，2004）。这些基于生物多样性的植物数据库的出现，极大地促进了相关研究和公众关注程度。国内主要开发了中国生物多样性信息系统（CBIS）、中国物种信息系统（CSIS）、中国科学数据库（CSDB）、中国科学院院生物局“生物物种与标本信息系统”和国家科技基础平台生物标本馆项目，以及不少的地方或专业的中小型库。中国生物多样性数据资源的建设通过以上项目的大力支持得到了长足发展。

1.1 国外同类著名数据库的建设情况与评价

国外尤其是欧美国家在植物数据库建设方面起步较早，现已建成多个著名的大型植物数据库。就提供名称信息查询而言，最著名的当数 species 2000，现在该数据库已经建立中国节点并收录“中国植物名录”（China Plant Catalogue，CNPC），提供植物基本名称信息查询，包括接受发表学名、异名、官方名（CNPC 还提供中文官方名）、分类信息、作者引证及作者信息、数据来源和文献信息、在线相关链接等信息，但 species 2000 毕竟是基于生物多样性保护而建设的，提供名称信息并不是其主要任务，所以 species 2000 也只能是提供植物名称基本信息的一个“植物名录”而已，不能进一步提供植物命名上更为深入细致的信息。相比较 species 2000 而言，国际植物名录（The International Plant Names Index，IPNI）是专业的植物名称数据库。IPNI 以 APNI，GCI 以及 KI 三大国际权威性数据源作为其后台数据库，使得其具有了得天独厚的先天优势，也被写入了《国际植物命名法规》作为推荐参考的国际标准植物名称查询系统。IPNI 的优势不仅体现在其权威性，还在于其专业性和全面性，除了能够提供大多数植物名称基础信息外，还提供了详细的作者信息、文献引证信息，并且能够通过多种方式查询植物名称信息。然而 IPNI 过于专业化，虽然在学名信息查询上优势明显，却不能够提供更多的其他名称信息，比如通用名，官方名等等。其它提供植物名称信息查询的还有 the Integrated Taxonomic Information System（ITIS），密苏里植物园的 TROPICOS® 等，他们提供的查询内容都大同小异。不过，国外数据库的最大优势，并不是权威性和海量数据，而是各个大型数据库之间拥有一套完善的交流分享机制，可以互相查询，互相整合，形成一个植物科学的强大网络知识体系。

1.2 国内同类数据库的建设情况与评价

国内早在上世纪 90 年代就已经开始了植物数据库的规划建设，至今已经有中国生物多样性信息系统（CBIS）、中国物种信息系统（CSIS）、中国科学数据库（CSDB）、中国科学院生物局“生物物种与标本信息系统”、中国科学院昆明植物研究所的《中国植物志》数据库、《中国种子植物》数据库和《云南高等植物电子词典》等大中型数据库，另外也出现了众多专业和地方性的中小型数据库，如景观植物信息查询系统（LPIIS）（韩成峰和张志国，2004），以及《湖北省植物志》数据库（董梅等，2005）等。这些数据库的建设与应用对我国植物科学，尤其是植物生物多样性的研究和保护起到了重要的推动作用。不过，相对而言，专门提供植物名称信息查

询的数据库在我国十分罕见，较为成熟的仅见中国科学院植物研究所的“中国植物名录(CNPC)”，而且还是 species 2000 国际合作项目的子项目。其他数据库只是顺带提供接受发表学名和中文官方名的查询，收录的数据太少以至于不能够满足用户需求；部分数据库缺乏维护，收录的数据“年代久远”而严重过时；某些数据库还大量设置访问权限，需要相当权限或者高额费用才能访问其核心数据。基于这些现状，我国急需要一个自主研发的，内容全面，信息权威，查询科学，实时更新，自由共享的新一代植物名称参考型数据库。

2　数据基础与建设内容

2.1　植物的名称数据组成分析

根据最新《国际植物命名法规》的相关规定，植物的名称包括接受发表学名、异名、保留名等等，而其中异名又分为分类学异名、命名学异名、基原异名等等。从单个的学名来说，对于种这一分类等级，完整的学名由属名、种加词、命名作者、来源文献组成，对于种以上的分类等级，则由名称、命名人和来源文献构成，对于种下等级，则还包括相应的标识、如变种（var.）、亚种（subsp.）等等，以及相应的加词、命名作者和文献来源。这些组成部分按照一定的规则排列得到一个完整的植物名称。一条完整的植物名称记录可以看作一条按一定规则生成的编码，亦可转换为相应的条形码，作为植物名称的唯一识别依据。

2.2　中国植物名称数据库的建设内容、数据组成与来源

中国植物名称数据库的目标是实现“中国植物电子名录”和“中国植物名称电子词典”以及“中国植物物种信息数据库查询索引”三大功能。基础数据库由“一库两名录”构成，即：一个名称数据库和两个彼此联系而又相互独立的《中国植物名录》（分别基于《中国植物志》和《Flora of China》）。所包含的内容包括植物的拉丁名称（接受发表学名，异名）、中文名（官方名，通用名，地方名，行业名等）、植物命名信息（接受发表学名详细，作者引证，文献引证，植物志索引），分类信息（科属，种及种下等级）四大块。另外还包含了一些为了实现上面“三大功能”和智能化检索所加入的其他预处理信息以及开发的小工具。根据用户需求，还加入了分布信息、特有标识和生活型等少量“非名称信息”。

中国植物名称数据库的数据分为基础数据和参考数据。基础数据是中国植物名称数据库的主要数据和核心数据，占绝大部分。作为参考型数据库，必须保证数据的正确与规范，所以除了专门制定相关数据标准规范外，基础数据全部来源于权威工具书《中国植物志》和《Flora of China》。由于志书本身记载不完善，有少部分没有记载的信息就要从其他资料或来源搜集，这部分就是参考数据。参考数据来源于其他经考察认为相对可靠的途径，按照事先制定的标准规范，经过考证后加入，由于无法完全确定其正确性，所以仅供参考。

3　建库策略实现

3.1　找准定位和用户对象

3.1.1　*专门针对特定的用户对象*　一个好的数据库服务系统，首先是一个受用户欢迎和好评的系统。所以要建设一个好的植物名称数据库，首先要搞清楚面对的是什么样的用户，了解用户需要什么。正如 species 2000 面向的是与生物多样性研究与保护相关的科学家，所以十分注重物种收录的丰富程度和地区性，USDA 面向美国的广大农业和园艺工作者和爱好者，所以专注于收录有用的美国资源植物，并提供详细的栽培养护，应用开发等信息（Singh，2004）。中国植物名称数据库面向的是植物学及相关学科的科研人员，他们不仅需要知道接受发表的学名和中文通用名，还要了解名称的来源、分类等级、中文俗名等等更为详尽的信息。所以中国植物名称数据库只提供接受发表的学名是远远不够的，还要提供相关的文献来源、植物的异名、中文通用名、俗名等等，全方位的满足植物学研究人员所需。

3.1.2　*根据用户需要建设，尽量全面提供信息*

在中国植物名称数据库中，按照惯例提供了常规的名称信息查询检索，如分类等级（科属），接受发表学名、异名、详细的学名信息（属名，种加词，作者信息，种下等级及作者信息）、名称来源文献、中文名（通用名，俗名，地方名，

行业名等)。在国内外已经建成的植物数据库中，提供的植物名称信息大体也是上述内容。

然而根据对用户的走访调查和用户反馈，发现用户查询植物名称的时候，不仅仅只关心植物的名称信息，还关心植物的特有性、分布、生活型等“非名称信息”，尽管中国植物物种信息数据库有专门的子库提供相关信息查询，但用户表示“不太方便”，根据用户这一需求，加入了一些“额外”信息，包括中国特有标识、分布(省，自治区，直辖市级)、简要的生活型（草本，木本等)，以方便用户所需。

根据用户需求和实际情况以及标准规范化的要求，还提供独立的可供下载查询的《中国植物名录》。

3.2 数据共享策略

中国植物名称数据库是一个参考型数据库，在不造成版权纠纷的前提下，原则上面向广大用户免费、自由地提供信息，在信息检索服务上，除了少数可能涉及版权问题的资源，绝大部分资源不再设置访问权限。

在数据库中添加了对应的流行通用植物编号代码等，如中国植物分类标准代码，仅由属名、种加词构成的“通用查询代码”，利用这些编号代码可以方便的在不同数据库中进行数据移植，实现数据共享。

中国植物名称数据库，应该是人人都可以自由分享信息的新一代植物名称数据库。

3.3 控制数据来源

3.3.1 *数据来源* 为了保证数据的正确性与准确性，必须有可靠的数据来源。《中国植物志》及其英文版是我国植物科学的权威资料，是四代植物学家的心血结晶，经过了历史和科学的验证。因此，中国植物名称数据库中的核心数据(接受发表学名，异名，名称详细信息，中文通用名，分类等级）全部来自于《中国植物志》及其英文版。这样可以有效规避因为数据来源不明而产生错误的风险。

3.3.2 *数据规范* 为了确保中国植物名称数据库所收录数据的准确性与严谨性。特别制定了《中国植物物种信息名称数据库标准规范》，以《国际植物命名法规》为基本原则，以《中国植物志》及其英文版为数据依据，结合参考型数据库建设的相关规范，对收录的植物名称进行规范化，标准化管理，杜绝张冠李戴和拼写错误，及时更新过时数据。

3.3.3 *数据完善* 植物科学的发展日新月异，中国植物名称数据库也应该是动态实时更新的，及时修正过时数据，比如接受发表学名因种种原因被废弃为异名，植物分类群地位变化等。在数据完善的策略上，引入目前流行的“open & free”模式，通过用户的参与，采用“普通用户提出＋管理员确认修改”和“权威专家直接修改”两种方式，结合传统的“管理员直接修改”方式，实现数据库的实时动态更新。

3.4 结构化数据

将植物的标准名称数据各个组成部分解析后分别作为独立字段储存，同时又可以组合成完整的植物名称。这样，看似不规则的植物名称，变成了一段由中文通用名、分类等级（科)、属名、种加词、种命名作者、种下等级类型、种下等级加词、种下等级命名作者、异名标记、文献来源构成的一串代码，它们相互独立又可以组合成串，这样大大提高了检索效率。

3.5 查询服务建设

判断一个数据库系统好坏的另一个标准就是看用户能否利用该系统容易快捷的检索到需要的信息，所以部署一个好的数据检索方式是保证中国植物名称数据库建设成功的重要前提条件。

在已建成的国内外数据库中，几乎都是采用传统的检索功能，智能化技术应用的相当少，中国植物名称数据库定位是新一代的植物数据库，所以在开发过程中应用了一些已有的技术和搜索策略，使之尽量实现智能化，最终是呈现给用户一个“会思考”的智能中国植物名称数据库。

3.5.1 *检索策略* 判断一个数据库系统好坏的另一个标准就是看用户能否利用该系统容易快捷的检索到需要的信息。所以部署一个好的数据检索方式是保证中国植物名称数据库建设成功的重要前提条件。

在结合对用户进行访问调查结果和以往其他成功的数据库产品的经验基础上，在中国植物名称数据库的检索方式部署上，采取了如下策略。

对于结构化的植物名称数据，既可以单独查询某个部分，也可以组合查询部分记录或整条完整记

录。这部分采用传统的关键词搜索，直接匹配。

对于植物的中文非官方名，如俗名、地方名、行业名等，即所谓“中文异名”。这些名称多而复杂，异物同名和同名异物现象非常普遍。这些数据通常不规则。但经过研究发现，这些“中文异名”的使用频率差异很大，有的使用率很高很普遍，甚至比植物志上记录的中文官方名还广为人知，而有的虽然存在这个名称，但使用很少，几近废弃。所以通过各种渠道对名称进行打分，如通过资料文献依据、查询频率、与用户互动等方式，按照打分高低排列，可以有效减少查出无关结果的概率。

3.5.2　*用户行为分析技术和智能代理技术*　通过记录用户的查询检索记录，对用户行为进行分析，将分析结果以访问 IP 为索引存档，可以了解用户的兴趣和检索习惯，当用户下次登录到系统时，利用智能代理技术，依据存档的用户行为分析结果，可以帮助用户快速方便地查询到自己需要的记录（夏勇，2009；郑玲和陈都，2007）。

3.6　科研工具开发

利用数据资源开发科研工具可以为科学研究提供极大的便利。例如基于生物多样性数据和地理信息系统开发的 BiodiversityMapping，可以用于地区生物多样性评价（赵海军和纪力强，2004）。考察国内外已经完成的植物名称数据库，几乎都是“电子植物名录”，“植物学名电子词典”，或者“电子版植物志索引”，没有进一步的开发利用。中国植物名称数据库作为参考型数据库，除了要实现传统的查询功能外，还要利用其本身的数据，开发一些工具，可以替代科研人员做一些简单的分析处理工作，提高科研的工作效率。

3.6.1　*文献知识索引系统*　通过事先对植物名称的解析和预处理，生成“查询代码”，包含了植物名称信息的主要部分，以此作为关键词，链接各大知识文献搜索引擎，进行文献查询。通过中国植物物种信息数据库的“内部编码 ID”链接到中国植物物种信息数据库的各个字库。通过“查询代码”或“国家标准代码”（国家质量监督检验检疫总局，1993）链接到其他植物数据库。上述结果返回后，可以给用户提供相当丰富的关于用户所查询相关植物的各类知识和文献资料。同时，尽量争取与国际大型植物数据库的合作共享，形成一个强大的植物知识网络系统。

3.6.2　*命名追溯系统*　中国植物名称数据库收录了大量的文献引证数据，这些文献引证数据记录了植物名称的发表和引用记录，包括了文献的名称缩写、页码、年代和其他说明。通过对文献引证数据进行分析，可以掌握一个植物名称的来龙去脉。中国植物名称数据库将分析文献引证的过程通过开发命名追溯系统的小工具，将极大地减轻植物命名追溯的工作量。

4　讨论

中国植物名称数据库是中国科学院科学数据库中国植物物种信息数据库的一部分，也是整个物种信息数据库的物种名录和查询索引，所以在建设过程中就一定要突出其“物种名录”和“索引”的功能。所以在建设过程中，就要特别注意收录名称的正确性和权威性，建立严格且有操作性的名称标准规范。在建设中采用的“名录独立，索引分开，编码联系”策略，就是一种既保证植物名录规范准确又体现索引功能的有效手段。由于数据来源本身可能存在多版本、重复记录、失误甚至错误，如果简单照搬中国植物志的名录索引作为《中国植物名录》，必然存在很大问题，所以建立独立的《中国植物名录》，以《中国植物志》名录为基础，通过多来源，多渠道的方法，反复订正，有效排除错误。同时，增加对应的内部及通用编码代号，同名称数据库与物种信息库其他部分以及其他数据库连接起来。

中国植物名称数据库的定位是参考型数据库，要求具有相当的规范性和权威性。采用《中国植物志》和《Flora of China》作为基础数据来源可以保证基础数据的规范性和权威性。然而在实际操作中，由于志书记录不完整的少量信息，需要后期进行补充，这部分数据的权威性就难以得到保证，所以只能作为参考数据看待。

植物的学名，包括接受发表学名和异名，是一种结构化的数据，而且不重复，可看作一串规则编码，认定容易。但植物的中文名，则是相当复杂，数量多而且重复率高，因此在中国植物名称数据库中，认定植物的中文官方名就显得十分不易。原则上以《中国植物志》记录为准，同时参考实际使用情况，可以保证大部分的中文官方

名认定是合理可靠的。

在数据检索方面，要达到真正意义上的智能搜索，在技术上现在还难以实现，所以我们需要通过其他技术手段使搜索尽可能的“智能化”，比如用户行为分析技术和智能代理技术。然而，记录用户的行为可能会构成隐私纠纷，还可能被杀毒软件误判为木马。解决的方式可以采用在页面告知用户，用户可以自行选择是否需要相关服务，并与网络安全公司合作，得到其安全软件认证，条件允许的话还可以向国家权威部门申请安全认证。

中国植物名称数据库作为中国植物物种信息数据库的起步，其成功建设可以为中国植物物种信息数据库其他子库的建设提供一个模式，同时为其他子库的建设奠定了名录和索引基础。随着计算机技术的发展，越来越多的新技术新方法将逐渐引入，数据库将越来越丰富化、智能化。以极其丰富的数据资源为基础，利用先进的计算机技术，比如云计算技术等“未来技术”，还可以建立一套有效的植物科学研究专家系统，为推动植物科学的进步革新提供部分动力。

〔参 考 文 献〕

中国科学院计算机网络信息中心科学数据中心，2009. 中国科学院数据应用环境建设与服务－参考型数据库建设规范 [M].

张丽兵译，2007. 国际植物命名法规（维也纳法规）[M]. 北京：科学出版社；圣路易斯：密苏里植物园出版社

国家质量监督检验检疫总局，1993. 中国植物分类与代码 [M]. GB/T 14467-1993

郑玲，陈都，2007. 用户行为分析在智能化搜索中的应用研究 [J]. 中国电力教育，**2**：386—387

赵海军，纪力强，2004. 生物多样性评价软件 BiodiversityMapping 的设计与实现 [J]. 生物多样性，**12**（5）：541—545

夏勇，2009. 网络信息检索与智能化搜索引擎 [J]. 计算机与网络，**14**：226—226，228

韩成峰，张志国，2004. 景观植物信息查询系统（LPIIS）的构建 [J]. 山东林业科技，**151**（2）：41—44

董梅，王健，中山诚宪等，2005.《湖北省植物志》数据库的设计与构建 [J]. 湖北林业科技，**1**：27—30

裴盛基，淮虎银，Alan Hamilton 等，2009. 植物资源保护 [M]. 北京：中国环境科学出版社

Singh G，2004. Plant Systematics：An Integrated Approach [M]. Science Publishers Ltd